Rental Property ROI Analysis

Sep 20, 2021 - Submitted by: Clifford Gong (clgong), Gen Ho (genho), Xinqian Zhai (xinqianz)

Motivation

The U.S. housing market has undergone a meteoric rise during the Covid pandemic. Redfin, one of the leading real estate brokerages, forecasted a record of $2.53 trillion worth of home sales in America in 2021, which is a 17% year-over-year gain that would mark the largest annual increase in percentage terms since 2013 [1]. While the housing market was suspended by the shelter-in-place order at the beginning of the pandemic, it has bounced back quickly with almost unstoppable momentum fueled by record low mortgage rates and persistent low inventories. The trend is expected to keep on.

However as home prices continue to go up, what about the rental market? In this project we study the before and since pandemic Residential Rental Property Return on Investment (ROI) for any substantial pattern change, and try to understand which factor contributes the most. We also try to identify which counties are more favorable to rental property investors . For rental property investors, we define them as investors who aim for steady cash flow (as opposed to those aiming primarily for property appreciation), and we assume they require mortgages for their purchases.

Overall, this study would like to answer the questions of 'when' and 'where'.

- For the question of 'when', we want to find out among Before Pandemic and Since Pandemic, which period is more favorable to rental property investors who aim for steady cash flow. This involves studying the ROI trend for rental properties, and how selling price, rent and mortgage rate influence the change of trend, if any.
- For the question of 'where', we want to discover which counties are more favorable to rental property investment. This involves studying the ROI trend for the housing market in different counties, plus the investigation of corresponding demographic data like population density and median household income, etc to identify possible patterns.

Data Source

Property Market Datasets

Home Values Data:		**Rental Data:**		**Property Tax Data:**	
A monthly frequency and county level Home Value Index(ZHVI) from 1996 to 2021		A monthly frequency and zip code level Observed Rental Index (ZORI) from 2014 to 2021		A county level average effective property taxes rate	
Key Features	RegionName, State, Home value from 2017-12 to 2021-08	Key Features	RegionName, State, MsaName, Rental value from 2018-01 to 2021-08	Key Features	County, Average Effective Property Tax Rate
Size	2,875 rows * 316 cols	Size	2,233 rows * 95 cols	Size	373 rows * 1 col
Format	CSV	Format	CSV	Format	CSV
Location	Click here	Location	Click here	Location	Click here
Access	Download	Access	Download	Access	Web Scraping

Supporting Datasets

Mortgage Rate Data:	
Fixed rate mortgage average in the U.S from 1971 to 2021	
Key Features	Date, Mortgage30us
Size	2,622 rows * 2 cols
Format	Excel
Location	Click here
Access	Download

Zip Code Data:	
All zip codes from country level to city level	
Key Features	Zip, Country, State, County
Size	42,632 rows * 15 cols
Format	CSV
Location	Click here
Access	Download

Home PPSF Data:	
State level price per square foot data data from 2012 to 2021	
Key Features	Region, price value from 2018 to 2021
Size	2068 rows * 65 col
Format	CSV
Location	Click here
Access	Download

Demographic Datasets

2018 Population Data:	
Annual Estimates of the Resident Population for Counties in 2018	
Key Features	California, Texas, New York, 2018
Size	373 rows * 1 cols
Format	Excel
Location	Click here
Access	Download

2021 Population Data:	
Population of counties in CA, NY, and TX in 2021	
Key Features	Name, 2021 Population
Size	374 rows * 4 cols
Format	CSV
Location	Click here
Access	Web Scraping

Other Demographic Data:	
Complete demographic comparison report of customized counties in 2021	
Key Features	Population density, Median age, Median household income, Vacant housing units
Size	596 rows * 9 cols
Format	CSV
Location	Click here
Access	Download

Data Manipulation

The focus of our data manipulation is to facilitate the calculation of ROI for each point in time. We define ROI in this study as the cash-on-cash return with the following formula [2]:

Cash-on-cash return = (gross rental income - expenses) / initial cash out of pocket

where

*gross rental income = monthly rent * 12*
expense = yearly mortgage payment + property tax + maintenance and management
initial cash out of pocket = down payment + closing cost + remodel cost

Monthly rent is retrieved from Rentals dataset. Yearly mortgage payment is calculated with selling price from Home Values dataset, and interest rate from Mortgage Rate dataset, assuming a 25% down payment. Property tax is retrieved from the Property Tax dataset. Maintenance and management is assumed at 10% of gross rental income. Closing cost is assumed to be 5% of selling price. Finally, the remodel cost is assumed to be $10,000.

In order to fulfill the above equation, we need to unify the keys, i.e. Date, County, across all datasets. Our target date range is from Jan-1-2018 to Jun-30-2021, and our counties are in California, New York, and Texas (explanation in Analysis and Visualization section) . This data manipulation work is done in the return_calculation.ipynb notebook.

Home Values

The Home Values dataset contains selling prices for each month from 1996-01-31 to 2021-07-31, across all counties in the U.S. After loading the csv file into Pandas, we filtered the data to our target date range and required states. The date value is at the end of each month, which needs to be transformed to the beginning of each month in order to join with other datasets. We used the datatime.striptime() to turn the date string into datetime type, followed by timedelta(days=1)) to add one more date to become the first day of next month, finally with datatime.strftime() to convert it back to string.

Rentals

The handling of Rentals dataset is similar to Home Values. However, there are three major problems with this dataset. Firstly, the record is at zip code level rather than county level. To overcome this, we found a Zip Code dataset, which contains state, county, and zip code. We joined these two datasets using dataframe.merge() with zip code as key, then used dataframe.groupby() to group the records at county level and take the median rent to represent each county. The second problem is there are null values. This is handled by dataframe.interpolate() which fills up the missing value with an average of two adjacent values. The third problem is we have found this dataset does not cover all counties. It only covers 29%, 8% and 20% for CA, TX and NY respectively. This literally limits the scope of our study.

Mortgage Rate

The Mortgage Rate data is in Excel format. To read this, we need to install the xlrd python package. Then the excel file can be read as a dataframe with dataframe.read_excel(). The Mortgage Rate records are created every Thursday. To align with our monthly resolution, we first extracted the string of year and month from the date column, then used dataframe.groupby() to group records with year and month, followed by picking the first rate. Finally, we formatted the year and month string to include the first day of a month. The rate column is percentage, and we divided it with 100 to transform it into decimal numbers.

Property Tax

Property Tax data were captured by web scraping, details in webscrap_PropertTaxRateData.ipynb notebook. We have individual csv files for the 3 states under investigation. After reading into dataframes, we vertically concatenated them using dataframe.append().

Returns

This is the core of this data manipulation process. We firstly got the date (index) and county (column) values lists from the Rentals dataframe. Then we used these two lists as keys to loop through the Home Values and Rentals dataframe, plus extracted appropriate values from Mortgage Rate and Property Tax dataframes. Once all the required variables were ready, we plugged them into the cash-on-cash return formula mentioned above to come up with the ROI value. Next, we transposed the Home Value, Rentals and Returns dataframes from wide format to long format using dataframe.melt() with date as the identifier variable. Finally, we joined these 3 dataframes using dataframe.merge() with data, county, and state as

key. Then we further merged with the Mortgage Rate dataframe with date as key to come up with the combined dataset for downstream analysis.

Population growth rate

Since there is no direct data on the population growth rate for a customized period (2018-2021), we need to generate the population growth rate by preparing the county level population of 2018 and 2021. For the 2018 Population dataset, we skipped the footer and header, and used only the county and 2018 columns to get the population and combined them together using append(). For the 2021 Population dataset, we defined a function to web scrape the table of 2021 population and only keep the name and 2021 population columns for each state and combine them together using append(). Finally, we merged the population18_all and population21_all on the county column using merge() and made a growth_rate column to hold the values of population growth rate, which was calculated by applying a simple operation using the 2018 and 2021 columns. The challenge was getting data from the web. We solved it by leveraging the requests library. We got the data using request.get(), then used read_html() to read the text we requested into a pandas dataframe.

Other demographic factors

We have individual csv files for demographic variables of the nominated counties and lower-ranked counties. We only kept columns with key demographic variables (mentioned in the data source), and transformed the rows (demographic variables) as columns using T(), and merged it with the population growth rate dataframe on the county column to make separate demographic data frames for each county category. Then we made an aggregated demographic dataframe for comparing the differences of the demographic variables between two county categories. We aggregated the demographic dataframe we created previously by mean and merged them together on the county column. Then derived a new column to hold the percentage differences for the demographic variables based on the existing columns.

For demographic related data, the data manipulation work can be found in get_pop_growth_rate.ipynb and demographic_analysis.ipynb notebooks.

Analysis and Visualization

Define the data scope for analysis

To begin with, we need to define the data scope for this analysis. Our target date range is from Jan-1-2018 to Jun-30-2021.

- We set the covid pandemic began date as Mar-1-2020, then define the pre-covid pandemic period as Jan-1-2018 to Feb-28-2020, while since-covid pandemic period is from Mar-1-2020 to Jun-30-2021.
- For location, we study the change of Price Per Square Foot (PPSF) across all states in the U.S. within our target date range. We pick the top, middle and bottom states, which are California, Taxes and New York respectively.

For details can refer to the find_candidate_state.ipynb notebook.

Has there been any change of ROI trend since pandemic?

A good first look at the data can be as simple as to eyeball your target variable on a line chart and see if there are any obvious changes in the data such as major up or down trends. Here we plot (fig.1) the median ROI for all counties in the nominated states for the period from Jan-1-2018 to Jun-1-2021 along with an annotation line at the start of the pandemic. Here we see that there is no noticeable change in trends that seem to be related to covid. An upward trend starts around Jan 2019, continues to rise smoothly through the start of the pandemic and doesn't begin to go down until Jan 2021.

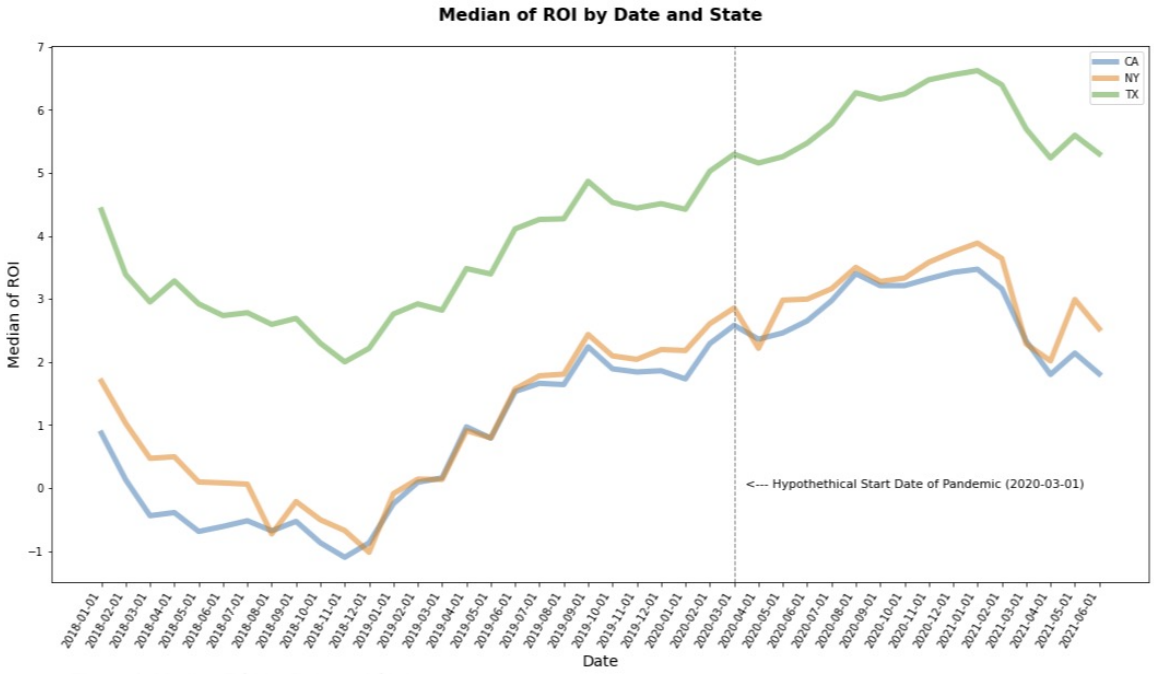

Figure 1: Median ROI by Date and State

It's a little anti-intuitive that the ROI trend wasn't affected by covid, because it's widely known that covid has had a devastating impact on the US and world economies. We next dive deeper into what affects ROI.

Among selling price, rent and mortgage rate, which factor impacts the ROI most?

We'd like to see which variable had the greatest impact on ROI to better understand what's going on in the data. Pearson's Correlation is a great tool to use here since we are dealing with continuous variables. However, first we should check to see how our data is distributed and if any groups exist.

(Marzban et al.) describes how important it is, when the data are not homogeneous, to distinguish between total and within-group correlation, because "by examining only the total correlation, one can miss the fact that the correlation within-group is very high" [3]. The authors go on to show an example of Simpson's paradox among others.

A SPLOM, or pair plot, can help us quickly see if there are any obvious groupings between our key variables. If there are, then we need to handle both total and within-group correlations. The pair plots revealed that ROI, Selling Price, Rent, Mortgage Rate, at the county level, contained groupings in all of the nominated states. In California state for example (fig.2), we can see obvious sub-groups at county level for ROI vs Selling Price and Selling Price vs Rent and other groupings as well:

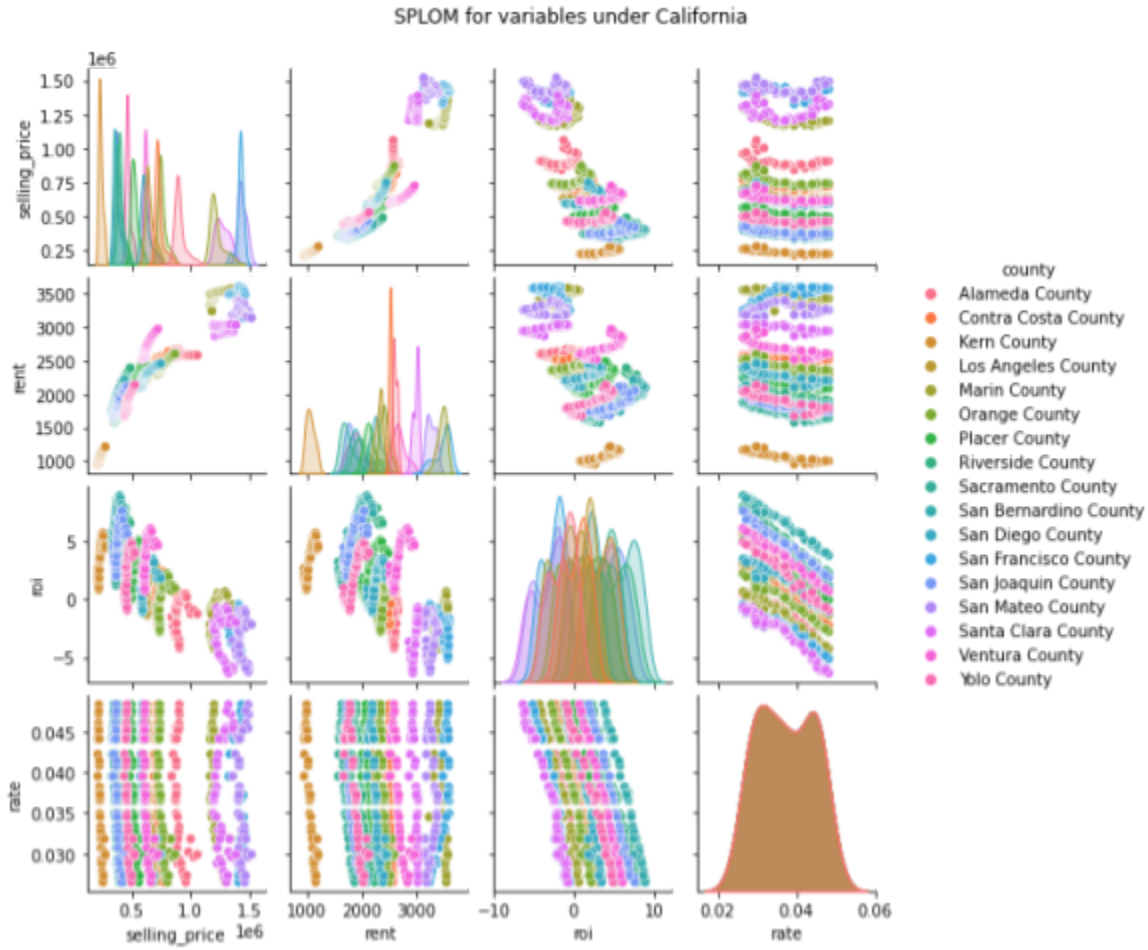

Figure 2: SPLOM for variables under California

We decided to take into account both the Total and Within Group correlation. We calculated the ***Total correlation*** by taking the Pearson Correlation between each of the variables with respect to all counties in that state. For the ***Within Group correlation*** we calculated the Pearson Correlation between each of the variables within each county first, and then took the average (mean) for each variable.

The results were quite different. Total correlation indicated that the greatest impact on ROI was Selling Price (CA: `-0.780`, NY: `-0.461`, TX: `-0.727`). However, Within Group correlation showed that, in fact, the correct variable with the greatest impact on ROI was Mortgage Rate (CA: `-0.984`, NY: `-0.943`, TX: `-0.963`):

Mean Correlation Coefficient Within Group vs Total Correlation Coefficient:

Features	Within Group	Total
---------- CA ----------		
ROI and Mortgages for CA Counties:	**-0.984**	-0.474
ROI and Selling Price for CA Counties:	0.502	**-0.780**
ROI and Rent for CA Counties:	0.646	-0.637
---------- NY ----------		
ROI and Mortgages for NY Counties:	**-0.943**	-0.442

```
ROI and Selling Price for NY Counties:          0.452  -0.461
ROI and Rent for NY Counties:                   0.528  -0.204
------------ TX -----------------------------------------
ROI and Mortgages for TX Counties:             -0.963  -0.355
ROI and Selling Price for TX Counties:          0.619  -0.727
ROI and Rent for TX Counties:                   0.752   0.224
```

Cross-checked with visualizations, we believe Within Group correlation has a better representation for the true relationships among variables.

Is there any change of correlation between Mortgage Rate and ROI before and since pandemic?

Knowing Mortgage Rate has the highest correlation with ROI, we further explore to see if there is any change in correlation before and since pandemic. We divided the combined dataset into 6 pieces by the pandemic start date followed by states (CA, NY, TX), then calculated the correlation coefficient for each county accordingly. The result (fig.3) shows almost all (46 out of 49, or 94%) counties exhibit a decrease of strength of correction since pandemic. We wanted to further investigate if the difference in correlation before and since pandemic is statistically significant. To answer this, we used *Fisher's z-transformation* for correlation coefficients pairs, then tested the null hypothesis that

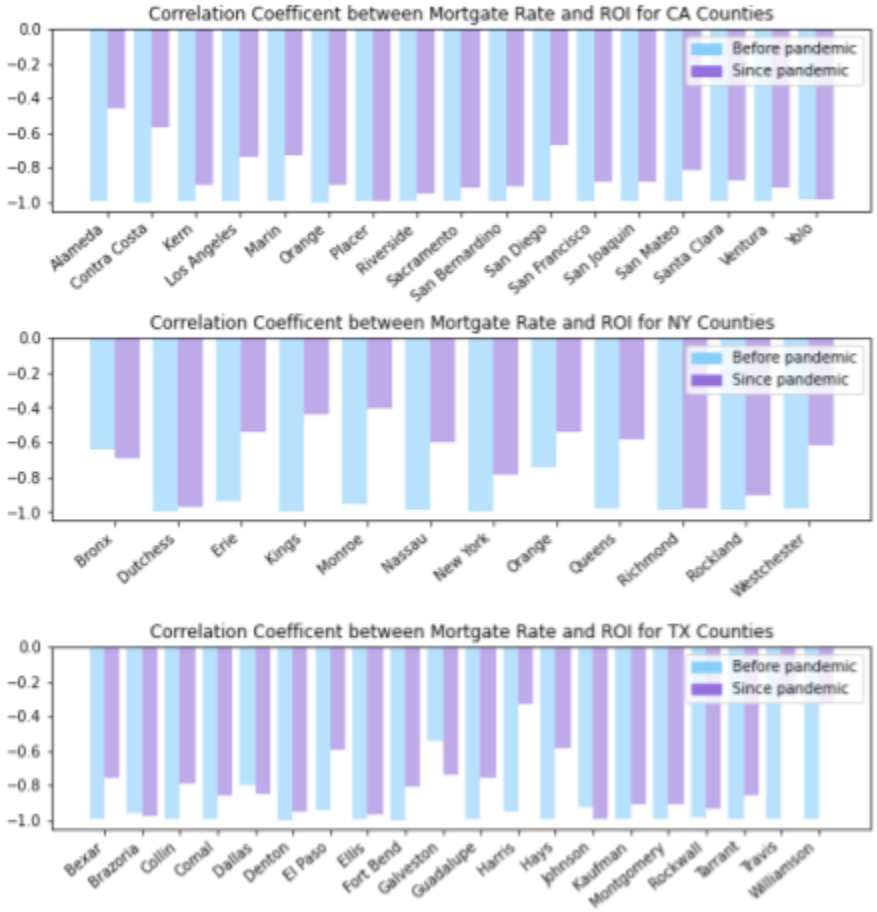

Figure 3: Comparison of Correlation Coefficient between Mortgage Rate and ROI for before and since pandemic

$r^{\,before\ pandemic} = r^{\,since\ pandemic}$ with $p < 0.05$ [4][5]. The result shows the two coefficients are significantly different for 43 out of 49 (88%) counties. Overall, this concludes that the strength of relationship between Mortgage Rate and ROI did change since pandemic. The whole correlation analysis can be found in notebook correlation_study.ipynb.

There is one limitation for this whole correlation analysis though - the Home Value and Rent data are smoothed by the data source provider (Zillow). If the data is smoothed before doing the correlation, some of the noises in the raw dataset will be average out, and may induce spurious correlations. When two time series are smoothed, their uncertainty would be reduced as data points are brought closer to their mean, which would potentially increase the correlation between them [6]. Should the data provider ever publish unsmoothed data, we can re-run the whole data pipeline to see if there is any significant difference.

Which county is more favorable to rental property investment?

The average mutual fund return for the past 15 years is around 7% [7]. With an ROI of 7% as the threshold, we nominated five counties that exceeded the median ROI of 7% across the period from January 2018 to June 2021:

Nominated counties	ROI(%)
El Paso County, TX	10.030
Kaufman County, TX	9.145
Johnson County, TX	8.910
Harris County, TX	7.980
Bronx County, NY	7.965

As we can see, 4 out of 5 nominated counties with the highest ROI are from Texas, one county from New York, and no county from California. So based on the data, Texas turns out to be a good place for rental property investors who aim for steady and high cash flow.

What are the demographic attributes for countries that are more favorable for rental property investment?

In order to see if there are some demographic differences between the nominated counties and the lower-ranked counties, we made a bar chart to show the percentage differences between them.

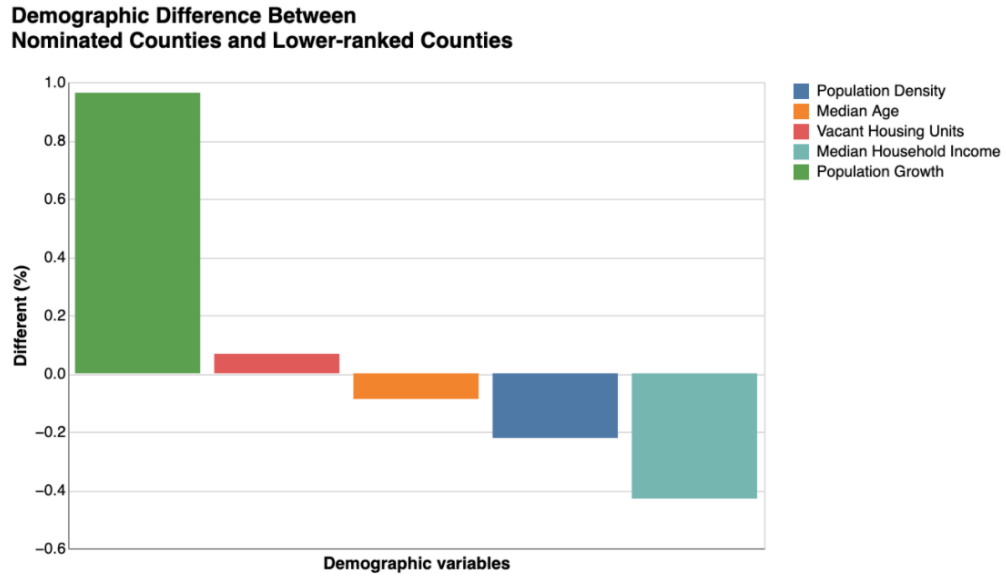

Figure 4: Demographic Differences Between Nominated Counties and Lower-ranked Counties

Compared with the lower-ranked counties we found that, on average, nominated counties have higher population growth rate and vacant housing units. They also had lower median age, population density, and median household income.

Among these factors, the average population growth rate of the nominated counties is almost twice that of the lower-ranked counties (fig.4). High positive population growth rate could be a good indicator of potential increases in demand for rental housing.

After looking at the demographic factors in an aggregated dimension, we further dived into each demographic factor at the county level. Here we focus on the average population growth since we showed it was the largest factor earlier (other factors are in the notebook).

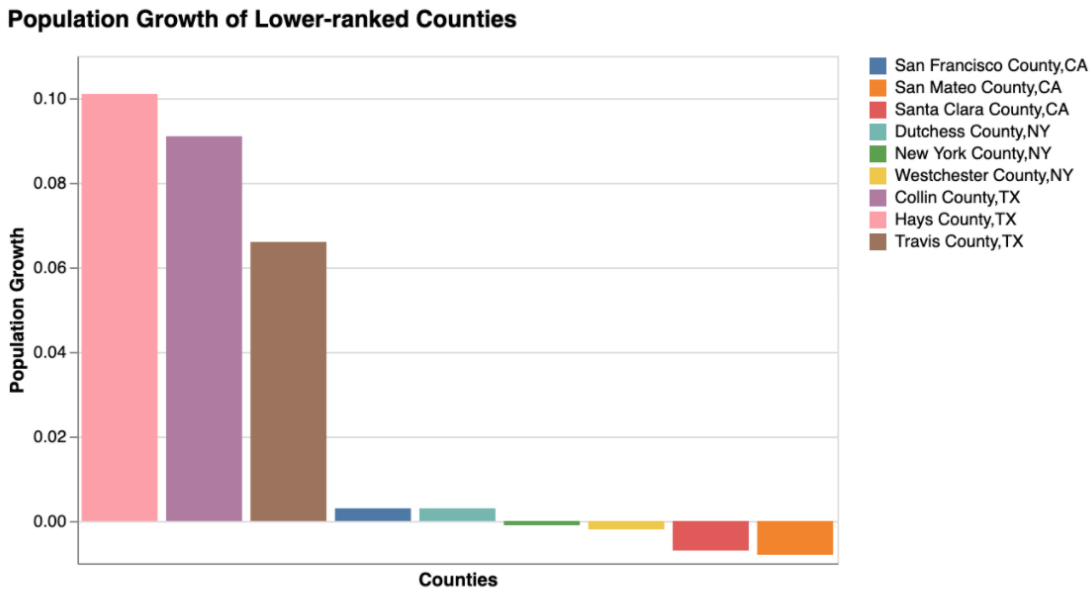

Figure 5: Population Growth of Lower-ranked Counties

We found the main contribution to the average population growth rate in the lower-ranked counties came from the counties in Texas: Hays county (10.1%), Collin county (9.1%), and Travis county (6.6%). The contributions from the counties in California and New York states are close to zero (0.3%) in two counties, and all others are negative.

In general, based on the differences, areas with higher population growth rate but lower population density, younger communities and relatively lower median household income, seem to be the attributes for more profitable counties for rental property investors.

The above demographic data study can be found in get_pop_growth_rate.ipynb and demographic_analysis.ipynb notebooks.

Conclusion

In this study, we found the ROI (cash-on-cash return) trend within the period of Jan 2018 to Jun 2021 was not affected by covid pandemic even though the housing market has undergone a meteoric rise since the pandemic (fig.1). Among house selling price, rent, and mortgage rate, mortgage rate has the highest

correlation with ROI. We observed the strength of correlation between ROI and mortgage rate is reduced when comparing before and since pandemic period (fig.3), and the difference is statistically significant. Among the candidate states, namely CA, NY and TX, Texas is the most favorable for rental property investors who aim for steady cash flow. Also, counties with higher population growth rate but lower population density, younger communities and relatively lower median household income are more favorable (fig.4,5).

Statement of Work

Clifford	Gen	Xinqian
• Web data scraping • Correlation Analysis and corresponding vis generation • Report writing	• Combine datasets for ROI calculation • Hypothesis testing for correlation comparison and corresponding vis generation • Report writing	• Web data scraping • Demographic data analysis and corresponding vis generation • Report writing

References

[1] Katz, L. (May 11, 2021). *Housing-Market Mayhem: U.S. Home Sales Likely to Hit Record High of $2.5 Trillion In 2021*. Redfin. https://www.redfin.com/news/record-home-sales-forecast-2021/. Accessed September 18, 2021

[2] Jahnke, T. (Updated June 11, 2021). *How to Calculate ROI on a Rental Property to Find Great Investments.* Roofstock. https://learn.roofstock.com/blog/calculate-roi-on-rental-property Accessed September 18, 2021

[3] Marzban, C., Illian, P. R., Morison, D., & Mourad, P. D. (Jan. 20, 2013). *Within-group and between-group correlation : Illustration on noninvasive estimation of intracranial pressure.* University of Washington. http://faculty.washington.edu/marzban//within_between_simple.pdf. Accessed September 17, 2021.

[4] Singer, P. (Oct 25, 2013) *Statistical Significance Tests on Correlation Coefficients.* Medium. tps://medium.com/@ph_singer/statistical-significance-tests-on-correlation-coefficients-b9397380be55. Accessed September 17, 2021.

[5] *Comparison of correlation coefficients.* (n.d.). MedCalc Software Ltd. https://www.medcalc.org/manual/comparison-of-correlation-coefficients.php. Accessed September 17, 2021.

[6] Briggs, W. (Feb 14, 2008) *Do not calculate correlations after smoothing data.* https://wmbriggs.com/post/86/?doing_wp_cron=1632081423.0117781162261962890625. Accessed September 19, 2021.

[7] Thune, K. (Updated May 18, 2021). *What Is the Average Mutual Fund Return?* The Balance. https://www.thebalance.com/what-is-the-average-mutual-fund-return-4773782. Accessed September 17, 2021.